



## Study of Different Sentimental Analysis Techniques: Survey

Sunny Kumar, Dr. Paramjeet Singh, Dr. Shaveta Rani

Computer Science Engineering & Technology & GZSCCET, Bathinda,  
Punjab, India

---

**Abstract**— *The growth of technology such as WWW, Social networking, Internet of things (IoT), electronic media etc. are responsible for the generation of vast amount of data in our daily routine by interacting with Internet world, Social networking media etc. across the world. It is very necessary to analyse and understand this unstructured datasets. So, Sentimental [14] analysis is one of the technologies which is used for determining whether the given piece of writing is positive, negative or neutral. It is also known as opinion mining. It means that it is deriving the attitude of a speaker. The main aim of this technology is to check or discover how different people think or feel about particular topic, celebrity or politicians by using text mining, NLP, linguistic computation [6] etc. This paper shows the brief review of different sentimental analysis techniques that are apply on the different type of datasets like product review data collected from Amazon.com, twitter datasets, Online Movies Review data etc. The most popular approach is knowledge based feature extraction which uses bag of words technique to explore the sentimental analysis of twitter using R performed on different data set domain.*

**Keywords**— *Opinion, Sentiment Analysis, text mining, NLP, Knowledge Based technique, SVM, Machine learning Algorithm, Naive Bayesian.*

---

### I. INTRODUCTION

From the past few decades, new form of communication and message exchange techniques, such as micro-blogging and text messaging [18] over social networking sites, have emerged and become ubiquitous day by day. Social media technologies spreads at a fast pace in these days. It comes into different forms like micro-blogging, blogs, social network, weblogs, internet forums etc. Micro blogging [11] web-sites are rich sources of data for text analytics, opinion mining and sentiment analysis. For example, millions of fans and viewers from all over the globe use Social Media to share their thoughts and emotions about the games, teams, celebrity and politicians, thus created massive amounts of data. So, decision is made to collect, analyse and visualize some of the data to look for interesting insights and correlations.

In today's area of technology, internet provides all type of information and consists of huge volume of information. Generally, most of the people share their views, opinions [15] and ideas over the internet by using social media sites in the form of textual data. It is freely and publicly [6] accessible over the internet and put a great impact in taking decisions about particular object, politicians or any celebrity. This information is in the form of reviews, comments, tweets, posts and news etc. Today, people preferred electronic media and internet generated reviews for taking any decisions.

Sentimental Analysis means to know about the real opinion of different people regarding specific product, company, movies, news, any object and their attributes. It includes [18] number of different branches like natural language processing, machine learning, text mining etc. By using these techniques and models, data is classified in the form of news articles, blogs, tweets, online reviews etc. into the positive, negative or neutral sentiments [14].

Twitter is a kind of micro-blogging social networking site consisting of billions of users in these days. In this paper, number of tweets is collected from twitter [4] server by using Twitter API on different keywords. The sentimental analysis function classifies all the tweets into different categories like positive, negative, or neutral [1] opinion. Therefore, twitter is a good platform to take decisions about the current topics in the news, popular politician, and celebrity or about any product.

R studio [12] environment offers number of packages to analyse data, graphics designing etc. However, R language is mostly used for statistical problems. But, when the size of tweets data is large it becomes impossible for the R studio to handle it. So, R studio environment must be integrate with some other systems to handle large [10] data efficiently.

In this paper, a Twitter dataset using Twitter API is constructed and explored properly, and R [12] studio coding is used to preprocess the Twitter data, using knowledge based methods. An available lexical resource is used and applied to the Twitter corpus (Dataset).

### II. LITERATURE REVIEW

In 2015, Xing fang and Justin Zhan uses product review data which is collected from Amazon.com for sentimental analysis. The main aim is to tackle the problem of sentiments polarity categorization which is one of the fundamental problem of sentiments analysis [5]. For this, both sentence-level and review level categorization are performed with promising outcomes. The main software used is scikit-learn, an open source learning software package in python.

The main classification models used are Naïve Bayesian, Random Forest and Support Vector Machine. They conclude that in sentence-level categorization when 200 feature vectors are formed on 200 manually labelled sentences, the classification model show same level of performance based on their scores. It shows that all three models performed good on testing data but, random forest model performs quite best all of them. In review level categorization 3 million level categorization vectors are formed. In this paper, it is observed that both SVM [5] model and Naïve Bayesian model are identical in terms of their performances and these model are superior to random forest model.

In 2014, **Shailendra Kumar Singh, Sanchita Paul and Dhananjay Kumar** focuses on the comparative study of different approaches for sentimental analysis using different data sets. The most popular approaches are bag of words and feature extraction to deal with sentimental analysis of opinion related with cars, music, electronic. It has been concluded that NLP [6] and pattern based techniques used for product review performed with accuracy level 85.6-100. In case of machine learning, accuracy level is 56.2-91.0. It has been seen that NLP and pattern based performed better than machine learning. They also give an idea of emerging trends of sentimental analysis is of social issues that is mostly popular twitter and you tube.

In 2014, **Geetikagautam and Divakaryadav** proposed an idea of sentimental analysis of twitter data using various machine learning approaches and semantic analysis. They proposed that twitter is an area where different peoples posted their views opinions in terms of tweets on social media. They contribute to sentimental analysis for customer review [18] classification model to analyse the information in terms of tweets that tweets data are highly unstructured and either positive negative or somewhere in between them. They first pre-process dataset after that, extracts the main adjective from dataset called feature vector and then applied machine learning based algorithms like Naïve Bayes, Maximum entropy and SVM along with the Semantic Orientation based WorldNet which extracts synonyms and similarity for the content feature. It has been seen that Naive Bayes gives better results than the maximum entropy and SVM is being subjected to unigram model gives a better results than using it alone. For more accuracy, semantic analyses WorldNet is followed up by the above procedure taking it to 89.9% from 88.2%.

In 2013, **Neethu M S and Rajasree R** uses machine learning techniques for twitter data for performing sentimental analysis. They proposed an idea that the sentimental analysis of twitter is quite difficult as compared to other sentimental analysis due to presence of slang words, and misspellings. They uses twitter posts about electronic products like mobiles, laptops etc. and uses knowledge based and machine learning approaches. They have used bag of words approaches [4] in symbolic techniques for sentimental analysis. In this, relationship between individual's words is not considered and whole of document is represented as a collection of words. It has also used machine learning techniques consisting of a training set and a test set for classification purpose. For this, they have used Naive Bayes \ (NB), Maximum Entropy (ME), and Support Vector Machines

(SVM) are used to classify reviews. They conclude that sentimental analysis of twitter is difficult to handle so they first pre-process all data and then feature vector is created by doing feature extraction.

Classification accuracy of the feature vector is tested using different classifiers like Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers. All these classifiers have almost similar accuracy for the new feature vector. This feature vector performs well for electronic products domain.

In 2012, **Federico Neri Carlo Aliprandi Federico Capeci Montserrat CuadrosTomas** have presented an idea of sentimental analysis using 1000 facebook posts about new casts, comparing the sentiments for rai- the Italian public broadcasting service - towards the emerging and more dynamic private company La7. They have studied results with observations[7] made by Osservatorio di pavia, which is an Italian institute of research specialized in media analysis at theoretical and empirical level, engaged in the analysis of political communication in the mass media. This study takes also in account the data provided by Auditel regarding newscast audience, correlating the analysis of Social Media, of Facebook in particular, with measurable data, available to public domain. Its results accurately reflect the reality as described by the Osservatorio di Pavia and Auditel, high lightening the importance of Facebook as a platform for online marketing. This study has been performed by a Knowledge Mining system used by some security sector-related government institutions and agencies in Italy to limit information overload in OSINT and Web Mining. The linguistic and semantic approaches implemented in this system enable the research, the analysis, the classification of great volumes of heterogeneous documents, helping documental analysts to cut through the information labyrinth, analysts to take account of complexity of public views, assigning automatically a sentiment polarity, rapidly accessing all the potential texts of interest.

In 2011, **Apoorv Agarwal, BoyiXie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau** have examined Sentimental Analysis on twitter data. They introduce POS-specific prior polarity features. They explore the use of a tree kernel to obviate the need for tedious feature engineering. They used three models- unigram model, a feature model, and a tree kernel based model. For the feature based model, some of the features are used that are proposed in past literature and have also proposed new features. For the tree kernel based model, a new tree representation for tweets is designed. A unigram model is used, previously shown to work well for sentiment analysis for Twitter data, as their baseline. Their experiments show that a unigram model is indeed a hard baseline achieving over 20% over the chance baseline for both classification tasks. The feature based model which uses only 100 features achieves similar accuracy as the unigram model that uses over 10,000 features. The tree kernel based model, outperforms both these models by a significant margin. Two kinds of models are investigated such as tree kernel and feature based models and demonstrates that both these models outperform the unigram baseline. For the feature-based approach, feature analysis is performed that reveals the most important features that combine the prior polarity of words and their parts-of-speech tags. This paper, tentatively conclude that sentiment analysis for Twitter data is not much different from sentiment analysis for other genres.

Twitter is an enriched form of tweets data used for sentimental analysis as sentiments are conveyed in the form of one or more sentence rather than paragraphs. Human generated status (tweets data) sentiments are not always obvious, many tweets contain ambiguous information (Agarwal et al.2011), slang words and misspelled data.

Another concept is that when the data generated from Twitter server is used which is present in the form of large amount of tweets convey no sentiment such as linking to a news article, can lead to difficulties in data gathering, training and testing. Parikh, Movassate (2009) [14] presented a Sentiment analysis process that provides a means of tracking opinions and attitudes on the web and also determines whether they are positively or negatively received by the public.

### III. SENTIMENT CLASSIFICATION

Shailendra Kumar Singh, Sanchita Paul and Dhananjay Kumar presented the Sentiment classification in various formats in different domains. It explores that, Sentiments classification is done into different contexts like- positive/negative, like/dislike, good/bad, buy product/don't buy, in favour/against, favourable/unfavourable, optimistic/pessimistic [6] etc. It can be done on different levels that are described below in detail:

1. Document Level
2. Sentence Level
3. Entity or Aspect Level.

In Document level, each document express some view, opinion, like web reviews or open ended questions in market surveys. Documents don't represent a [2] single point of view, a single opinion. Document consists of multiple opinions representing several closely related but nuanced positions. Whole document can be reduced and classified into positive or negative class. Sentence level classifies each sentence into positive, negative or neutral class. Entity level product features are identified from the source data.

Sentiment analysis is usually conducted between two levels; a coarse level and a fine level. Coarse level sentiment analysis deals with determining the sentiment of an entire document and Fine level deals with attribute level sentiment analysis.

As per technical perspective there are two main approaches for sentimental analysis. They are given below:-

#### A. Symbolic Techniques

In July 2013, Neethu M S and Rajasree R proposed that Symbolic techniques also known as knowledge based approach [2]. In this technique, available lexical resources are used. In this sentiment analysis approach, bag-of-words approach is used. The BOW model focuses on the words list, or says string of words, it cannot check the context of the sentence. This model contains a list of words that have own value when found in the given text. This model totally focuses on the words and take care nothing about the language fundamentals.

The difficulty in using a Knowledge base approach is that it requires a large lexical database. This has become harder and harder to provide as the language of social networks is so trend dependent and changeable that lexicon datasets cannot keep up. Therefore, Knowledge based approaches to sentiment analysis are not as popular as they are used to be. Symbolic Techniques.

#### B. Machine Learning Techniques

In contrast to Knowledge based approaches, Machine Learning techniques are not using any lexicon resources list , instead a training set and a test set is used in order to [4] classify them. Training set contains input vectors and corresponding class labels for training the network. After that, test set is used to validate the given model by checking the class labels to unknown feature vectors. There are different machine learning techniques like SVM, maximum entropy [6] and Naïve Bayes etc. This allows the algorithm to remain dynamic in the face of ever changing social network language lexicons. In this methodology, a classification model is developed using a training set, which tries to classify the input feature vectors into corresponding class labels. Use the results from the knowledge based techniques and those of the machine learning techniques to ensure a thorough analysis of the dataset.

### IV. TEXT MINING PROCESS

Geetikagautam and Divakaryadav have explained the text mining process which shows the different steps that are used for sentimental analysis. These steps are shown in figure 1:

1. **Data Crawling from Twitter:** As no datasets are available on the internet of twitter site for analysis. A new dataset is created by collecting tweets over twitter server. Tweets were automatically collected using Twitter API. This is the first step in data mining process on Twitter.
2. **Data pre-processing:** Pre-processing step is used for removal of punctuation marks, correctors and digits, slagan words, noisy, inconsistent and incomplete data etc. as well as changing tweets text data to lower case, splitting sentences to words that are compared with the list of positive [18]and negative words list. Using R, the structure of a tweets data is studied properly and pre-processing of the tweets data is performed and eliminates all unwanted information [16].
3. **Sentiment Analysis of the Tweets:** The purpose of this step is to build an algorithm or function that can accurately perform sentiment analysis and opinion mining on the given data. As words are compared with given words list the matched term would be returned as a true or false which will be treated as a 1 or 0 by the sum function. The final score for each tweet will be the number of positive words minus the number of negative

words. If the score is higher than 0, the tweet will be regarded as positive. If the tweet score is lower than 0 the tweet will be regarded as negative opinion.

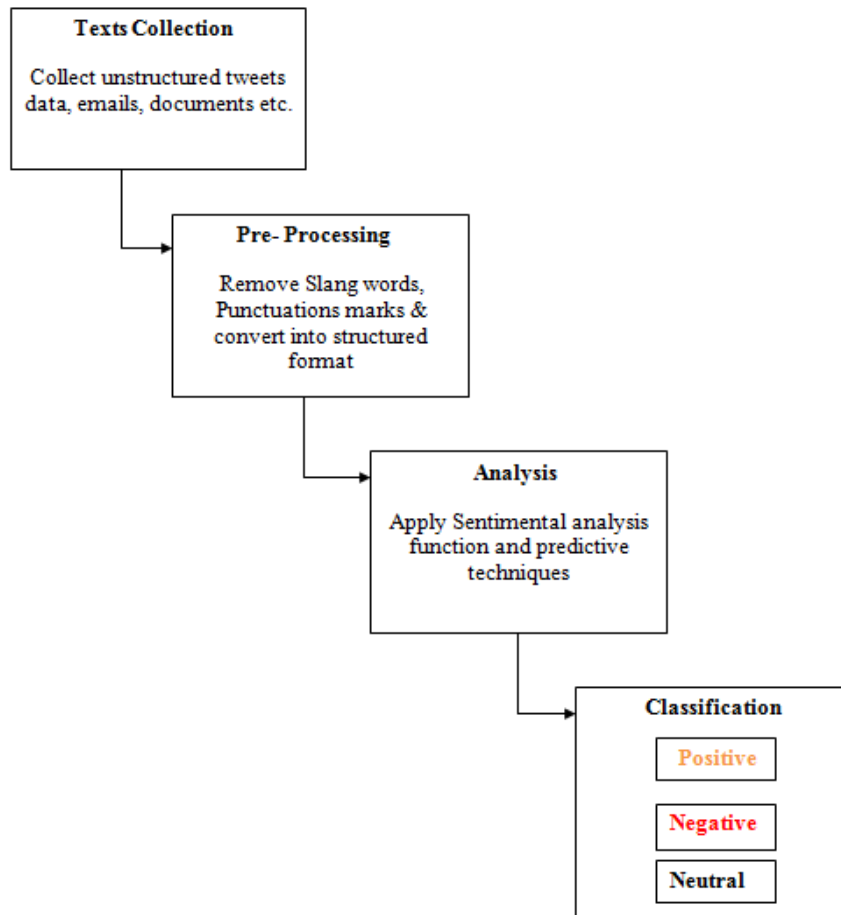


Figure 1: Steps in Text Mining Process

4. **Data Visualizations:** After collecting and analysing Twitter data, it is continued with some notions on data visualization using R-programming.

## V. RESEARCH WORK DONE USING DIFFERENT IMPLEMENTATION TECHNIQUES AND RESULTS

Anne Hennessy Student of National College of Ireland proposed Sentiment Analysis of Twitter Using Knowledge based and Machine Learning Techniques. The paper choose to analysis tweets with the hash tag #ConchitaWirst and #Eurovision2014. This paper decides to investigate the tweets related to wurst and the Eurovision song context 2014. The various steps are shown in below:-

### A. Knowledge Based Methodologies Algorithm Using Twitter Dataset

1. **Creating a twitter application-** First step in twitter sentimental analysis is to create a twitter application. This application allows to connect with the twitter server for crawling the data by using Twitter API.
2. **Installing R packages-** The number of packages is installed in R studio environment for connecting R console to twitter to apply commands and it gives proper results. These packages are- twitter, ROAuth, plyr, Stringr, ggplot2 etc.
3. **Handshaking-** Next step is to access the Twitter API. This step includes the script code to perform handshaking using the ConsumerKey and ConsumerSecret of the application.
4. **Saving Tweets-**After handshake is performed data is crawled related to the preference The function “search twitter” is used to download tweets from twitter server.
5. **Sentimental Function-** After collecting dataset, sentimental function is applied for analysis of data. This function is used to convert tweets data into some useful needy information. This step is to check the sentiments of different tweets of every user on the basis of positive score, negative score or in between them. The key task is to use the stored documents that contain positive and negative words list. The function uses this List and check the similarity with the words that the user uses in their sentences. Thus, it is helpful to show the polarity of the sentiment for the users. The sentiment function calculates score for each individual tweet. It first calculate the positive score by comparing words with the negative words list and then calculate negative score by comparing words with negative words list. The final score is calculated as  $\text{score} = \text{positive score} - \text{negative score}$ .

```
Source on Save Run
1 #####
2 # Chunk - 1 - Authenticate with twitter API
3 #####
4
5 install.packages("twitter",dependencies=T)
6 install.packages("plyr")
7 install.packages("stringr")
8 install.packages("ggplot2")
9
10
11 library(twitter)
12 library(ROAuth)
13 library(plyr)
14 library(stringr)
15 library(ggplot2)
16 library(streamR)
17
```

Figure 2: Loading of R packages

```
> # Accessing the twitter API
> requestURL <- "https://api.twitter.com/oauth/request_token"
> accessURL = "http://api.twitter.com/oauth/access_token"
> authURL = "http://api.twitter.com/oauth/authorize"
> consumerKey = "9St03woVMawP5VhIMsHHw"
> consumerSecret = "fDgjbFIOZCFThzhcAx5S7szX1xHrvixR0Eg18cUw"
> Cred <- OAuthFactory$new(consumerKey=consumerKey,
+                           consumerSecret=consumerSecret,
+                           requestURL=requestURL,
+                           accessURL=accessURL,
+                           authURL=authURL)
> Cred$handshake(cainfo = system.file("CurlSSL", "cacert.pem", package = "Rcurl") )
To enable the connection, please direct your web browser to:
http://api.twitter.com/oauth/authorize?oauth_token=dHwEGdxbj093sG0tvjYVTONQrkjU3DuCxcC1YQyc
When complete, record the PIN given to you and provide it here:
```

Figure 3: Authentication and Handshaking with Twitter

### B. Machine Learning Algorithm Implementation Using Different Datasets

Xing fang and Justin Zhan have collected dataset of product reviews from Amazon.com. From February to April 2014, they collected, in total, over 5.1 million of product reviews [5] in which the products belong to 4 major categories: beauty, book, electronic, and home. Those online reviews were posted by over 3.2 million of reviewers (customers) towards 20,062 products.

The process followed by algorithm is described below:

1. The main aim is to tackle the problem of sentiment polarity categorization which is the main problem of sentiments analysis.
2. The data used in this paper are online product reviews collected from Amazon.com.
3. Both sentence-level and review level categorization experiments are performed with promising outcomes.

Neethu M S and Rajasree R have implemented sentimental analysis using electronic products dataset domain. As there is no standard twitter dataset is [4] available for electronic products domain, a new dataset is created by collecting tweets over a period of time ranging from April 2013 to May 2013. Tweets are collected using Twitter API and they are manually annotated as positive or negative. A dataset is created by taking 600 positive tweets and 600 negative tweets.

In 2012, Federico Neri Carlo Aliprandi Federico Capeci Montserrat Cuadros Tomas have presented an idea of sentimental analysis. They have used 1000 facebook [7] posts about new casts. , by focus crawling of Facebook. Only the textual contributions related to La7 and Rai1 news programs have been semantically analysed and indexed.

```
# sometimes a list() is one level of hierarchy too much
words = unlist(word.list)

# compare our words to the dictionaries of positive & negative terms
pos.matches = match(words, pos.words)
neg.matches = match(words, neg.words)

# match() returns the position of the matched term or NA
# we just want a TRUE/FALSE:
pos.matches = !is.na(pos.matches)
neg.matches = !is.na(neg.matches)

# and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum()
score = sum(pos.matches) - sum(neg.matches)

return(score)
}, pos.words, neg.words, .progress=.progress )
scores.df = data.frame(score=scores, text=sentences)
return(scores.df)
}
```

Figure 4: Sentimental Analysis Function

## VI. RESULTS

1. Performance of each classification model is estimated base on its averaged F1-score whose formula is given below where  $P_i$  is the precision of the  $i$ th class,  $R_i$  is the recall of the  $i$ th class, and  $n$  is the number of classes.  $P_i$  and  $R_i$  are evaluated using 10-fold cross validation.

$$F1_{avg} = \frac{\sum_{i=1}^n \frac{2 \times P_i \times R_i}{P_i + R_i}}{n}$$

**Xing Fang and Justin Zhan** have used three classifications models [5] such as: Naïve Bayesian, Random Forest and Support Vector Machine. They concluded that in sentence-level categorization when 200 feature vectors are formed on 200 manually labelled sentences the classification model and show the same level of performance based on their scores as shown in figure 5. It shows that all three models performed good on testing data but random forest model performs quite best all of them.

2. **Shailendra Kumar Singh, Sanchita Paul<sup>2</sup> and Dhnanjay Kumar** have presented the comparative study of different analysis techniques and showed that NLP and Pattern [6] based techniques are used for the product review and performed with 85.6 – 100 accuracy. While in the case of machine learning accuracy range varies between 56.2 and 91.0. So, it is true that NLP and Pattern based techniques perform better than the Machine Learning for sentiment analysis of product review.
3. **Geetikagautam and Divakaryadav** give the results in tabular form as shown in figures. Table 1, Table 2 show the performance measures of Naive Bayes, maximum entropy and support vector machine based classifiers respectively in terms of precision and[18] recall. Similarly, Table 3 shows the performance of the classifiers terms of accuracy.

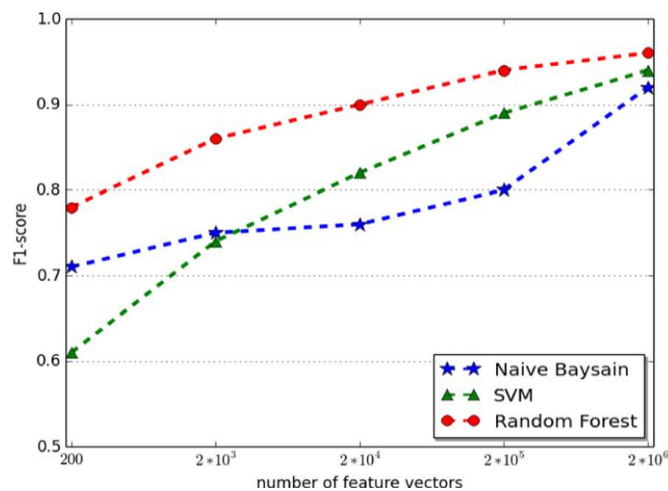


Figure 5: F1 scores of sentence-level categorization.

Table 1: Accuracy Comparison

Methods	Accuracy
Naïve Bayes	88.2
Maximum entropy	83.8
Support vector Machine	85.5
Semantic analysis(WordNet)	89.9

Table 2: Naive Bayesian and Maximum Entropy measurements

	Naive Bayesian Measurements	Maximum Entropy Measurements
Positive Recall	91.2	86.1
Negative Recall	85.4	80.0
Positive Precision	49.3	40.4
Negative Precision	39.3	33.6

4. **Neethu M S and RajasreeR** proposed that when twitter API is used to collect tweets and dataset of 1200 tweets are made of electronic products. It has [4] been seen that Naive Bayes has better precision as compared with other three classifiers, but slightly lower accuracy and recall. SVM, Maximum Entropy Classifier and Ensemble classifiers have similar accuracy, precision and recall. They obtained an accuracy of 90% whereas Naive Bayes has 89.5%. This shows that the quality of the feature vector is selected for the product domain. This feature vector aids in better sentiment analysis despite of the classifier selected.



Table 3: Accuracy level

Algorithm Name	Accuracy level
SVM , Maximum Entropy	90.0%
Naïve Bayes	89.5%

## VII. FUTURE SCOPE

Social media networking has been growing day by day so it is very necessary for various organizations to analyse customer behaviour or attitude of particular product or a celebrity, politicians. So, sentimental analysis has been introduced. Text mining and [9] sentimental analysis helps to derive valuable business insights. It is little bit complex task in case of twitter as tweets consist of short hands as online review were written in more clear way. Several classifications models and different classifications for various sizes of datasets are studied to address the topic.

It is observed that R language for sentimental analysis on a small dataset gives good results but sometimes when large dataset is used its performance degrades. So, in that cases it must be integrate with some other tools so that large dataset could be analyse using R. It is observed that twitter generates massive amount of data say size in TB's. So, it means Big Data?? To handle this large amount of data it must be combine with big data tools so that sentimental analysis can be easily and efficiently performed. It has been seen that Big data is an growing area in the field of computer science technology and it is an important application of sentimental analysis.

## VIII. CONCLUSIONS

In this paper, different classification techniques of sentimental analysis are studied. Knowledge based methodologies and different Machine Learning algorithm is used in order to give a thorough examination of the various tweets which were extracted from Twitter, face book or any other media. It has been explored that the bag of words approach uses a publicly available sentiment lexicon resources list, which consists of around 5500 words. To compare a positive and negative opinion words or sentiment words for English bag of words approach was used to separate the tweets. Sentimental Analysis of twitter is quite difficult as compared to other general sentimental analysis due to presence of white spaces, slang words, misspelling in tweets, short length, punctuations and numbers had to be confronted in the pre-processing stage. To further alleviate these issues, twitter specific features were extracted and added to the feature vector after proper pre-processing. From this paper, it is concluded that R language is a suitable tool for small dataset real time data analysis. It is also concluded that when size of dataset increases, physical memory exceeds and gives poor results. Therefore, it must we integrate with some other tools for better performance. It has also been observed that Machine Learning Algorithms such as- SVM, Maximum Entropy, Naive Bayesian give good results when they are performed on the different datasets.

## REFERENCES

- [1] M. Karamibekr and Ali A. Ghorbani, "Sentiment Analysis of Social Issues", International Conference on Social Informatics (IEEE), (2012).
- [2] R. Parikh and M. Movassate, "Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009
- [3] <http://cran.rproject.org/web/packages/ffbase/ffbase.pdf>
- [4] Neethu M S and Rajasree R," Sentiment Analysis in Twitter using Machine Learning Techniques" fourth ICCCNT international conference on July 4 - 6, 2013, Tiruchengode, India
- [5] Xing Fang and justin Zhan "Sentiment analysis using product review data." *Fang and Zhan journal of big data*, May, 2015.
- [6] Shailendra Kumar Singh, Sanchita Paul2 and Dhananjay Kumar," Sentiment Analysis Approaches on Different Data set Domain: Survey," Vol. 7 5 No.,pp.39-43,2014
- [7] Federico Neri Carlo Aliprandi Federico Capeci Montserrat Cuadros Tomas," Sentiment Analysis on Social Media," *International Conference on Advances in Social Networks Analysis and Mining*, june,2012
- [8] <http://cran.rproject.org/web/packages/RODBC/RODBC.pdf>
- [9] Liu B (2010) Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, Second Edition.
- [10] Taylor and Francis Group, BocaR. Feldman," Techniques and Applications for Sentiment Analysis," Communications of the ACM, Vol. 56 No. 4, pp. 82-89, 2013.
- [11] Z. Niu, Z. Yin and X. Kong, ' Sentiment classification for microblog by machine learning,' Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, vol. 286289, 2012.
- [12] Revolution Analytics , "Advanced Big Data Analytics with R and Hadoop,"
- [13] <http://www.revolutionanalytics.com/sites/default/files/r-andhadoop-big-data-analytics.pdf>, 2015, white Paper.
- [14] Mejova, Y. (2009). Sentiment Analysis: An Overview. Comprehensive exam paper, available on <http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf> [2010-02-03].

- [15] Pang and L. Lee, Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval*, 2008, pages 1-135.
- [16] <https://www.scikit-learn.org/stable/modules/preprocessing.html>
- [17] <http://www.r-bloggers.com/sentiment-analysis-on-donald-trump-using-r-and-tableau>
- [18] GeetikaGautam and Divakaryadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis," July, 2014. M. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.
- [19] <http://blog.dataweave.in/post/96618078833/building-a-twitter-sentiment-analysis-app-using-r>
- [20] Prabhat Kumar, BerkinOzisyilmaz, Wei-Keng Liao, GokhanMemik and AlokChoudhary, "High Performance Data Mining Using R on Heterogeneous Platforms," *IEEE International Parallel & Distributed Processing Symposium*, May, 2011
- [21] <http://www.r-bloggers.com/twitter-sentiment-analysis-with-r/>